

Method, Program Product and Apparatus for Discovering
Functionally Similar Gene Expression Profiles

5

ABSTRACT OF THE DISCLOSURE

Genes to be compared are listed by their gene expression
profiles and processed with a similar sequences algorithm that
10 is a time and intensity invariant correlation function to obtain
a data set of gene expression pairs and a match fraction for
each pair. A threshold match fraction is chosen and a null set
is created to hold indices of genes accounted for. Genes are
then assigned to clusters by match fraction value if they have a
15 match fraction greater than the threshold. Genes are then
removed from clusters if they are represented in more than one
cluster by removing a first gene from a cluster when another
cluster has another gene with a higher match fraction with the
first gene. When the difference between maximum match fraction
20 values for pairs including a first gene in a first cluster and
the first gene a second cluster is small, the first gene may be
removed from the first cluster even when another gene in the
first cluster has a higher match fraction with the first gene
than the first gene has with a third gene in a second cluster.
25 This occurs when the number of similar subsequences for the pair
including the first gene in the first cluster is higher than the
number of similar subsequences for the pair including the first
gene in the second cluster.